Micro-targeting and Electorate Segmentation:
*Data Mining the American National Election Study**

Gregg R. Murray
Department of Political Science
SUNY Brockport
Brockport, NY 14420
gmurray@brockport.edu
585-395-5676

Anthony Scime
Department of Computer Science
SUNY Brockport
Brockport, NY 14420
ascime@brockport.edu
585-395-2323

*This manuscript is a post-print. The final, formatted version is available from Taylor & Francis at http://www.tandfonline.com/doi/abs/10.1080/15377857.2010.497732.

Micro-targeting and Electorate Segmentation:
Data Mining the American National Election Study

ABSTRACT. Business marketers widely use “data mining” for segmenting and targeting markets. To assess data mining for use by political marketers, we mine the 1948-2004 American National Elections Study data file to identify a small number of variables and rules that can be used to predict individual voting behavior, including abstention, with the intent of segmenting the electorate in useful and meaningful ways. The resulting decision tree correctly predicts vote choice with 66% accuracy, a success rate that compares favorably with other predictive methods. More importantly, the process provides rules that identify segments of voters based on their predicted vote choice, with the vote choice of some segments predictable with up to 87% success. These results suggest that the data mining methodology may increase efficiency for political campaigns, but they also suggest that, from a democratic theory perspective, overall participation may be improved by communicating more effective messages that better inform intended voters and that motivate individuals to vote who otherwise may abstain.

"Bourbon drinkers are more likely to be Republicans; gin is a Democratic drink. Military history buffs are likely to be social conservatives. Volvos are preferred by Democrats; Ford and Chevy owners are more likely Republican. Phone customers who have call waiting lean heavily Republican." -- Hamburger and Wallsten (2005)

Technology continues to transform campaigns and political marketing. The emergence of fax, email, the Internet, and a variety of media and computer technologies has contributed to the rise of candidate- and consultant-centered campaigns (Shea and Burton 2006). The new technologies provide campaigns with direct access to voters and permit narrowly crafted messages to be delivered to very specific segments of voters. In an effort to take advantage of these tools, both major parties collect conventional information from publicly available records on registered voters such as party registration, voting history, and driver's licenses (Hamburger and Wallsten 2005). But the parties and strategists increasingly are enhancing these records with in-depth consumer data gathered from varied sources such as supermarkets, magazine subscriptions, charitable contributions, online book vendors, drugstores, automobile dealerships, and issue surveys (Edsall 2006; Hamburger and Wallsten 2005). Campaign experts have concluded that analyzing such detailed information, often using emerging and increasingly sophisticated techniques, may uncover previously unknown yet important relationships that can be used to identify more precisely potential supporters and, therefore, to micro-target more persuasive messages to them. One such emerging technique used for micro-targeting is data mining.

Of course, micro-targeting requires extensive and expensive data and some means to extract useful information from that data. Many supermarkets, for example, allocate substantial resources to construct databases that rigorously track the thousands of purchases made by their customers. They later analyze the data to identify buying patterns that may be exploited to increase future sales through techniques such as coupons and product placement. Among the memorable relationships found by this process is an actionable association between buying diapers and beer on Thursdays (Witten and Frank 2005). These types of relationships are uncovered when interested parties use data mining, which is a data analysis methodology that employs specialized algorithms to extract information from extensive data sets. For example, a leading telecommunications company acquired new customers and increased expenditures by those customers using data mining to identify client needs and trends related to characteristics such as number of family members, average family age, and geographic area (McCarthy 1997). And the power of these techniques is not being ignored by the political parties. Recent reports suggest that Democratic campaign operatives, following their Republican counterparts, intend to allocate substantial resources to creating and mining databases of existing and potential supporters (Edsall 2006). This results from a belief that enhancing conventional voter information such as place of residence with data such as type of restaurant frequented and genre of magazine read can produce a more accurate and refined prediction of a segment of voters' support for a candidate or position on a set of issues.

In this research, we propose to undertake a data mining analysis to evaluate the utility of data mining methodologies in the context of political marketing. In particular, we attempt to identify a small number of variables and rules that can be used to predict individual voting behavior, including abstention, with the intent of segmenting the electorate in useful and meaningful ways. In lieu of the proprietary databases that are often used for these analyses, we evaluate the American National Election Study (ANES 2005), a publicly funded, long-term series of public opinion surveys designed to produce research-quality data on voting, public attitudes, and political participation. The ANES has been conducted nearly every federal election year since 1948 and, in the process, has interviewed more than 47,000 respondents and resulted in more than 900 survey items designed to assess individuals' attitudes about national elections and their outcomes.

As such, this paper proceeds as follows. The next section introduces the basics of the data mining methodology. We review the types of data mining, the data mining process, and decision tree construction, which is the data mining process employed in this research. The following section describes the ANES data and the characteristics of that data set that are pertinent to and make it appropriate for data mining analysis. Then we review the results, which indicate that, indeed, the process yields a small number of successful predictors of voting behavior—i.e., voting for the Republican, Democratic, a major-party or other presidential candidate or, importantly, abstention—that can be translated into a set of rules used to identify specific segments of voters. In the end, we conclude that data mining outputs may have important implications for political campaigns and, more broadly, for those concerned with democratic participation. We finish with a brief conclusion and considerations for future research.

What is Data Mining?

Data mining is a process of inductively analyzing data to find interesting patterns and previously unknown relationships in the data. The data mining process involves both human and software resources (Hofmann and Tierney 2003; Scime and Murray, in press). Data mining is used not only to predict the outcome of a future event but also to provide knowledge about the structure and interrelationships among data. The data mining process identifies relationships that are expressed as classification rules, association rules, and decision trees. Classification and association mining algorithms are used to create models that describe existing data and relationships within the data. Both methodologies create rules used to analyze new data and predict future outcomes. Classification analysis constructs a decision tree, which is an analytical technique that is both explanatory and predictive (Osei-Bryson 2004) and, as a result, has a long history in both marketing and political science. Rules are derived from decision trees. The decision tree and resulting classification rules and, more importantly, the segments of the electorate they imply are the focus of this research.

Classification algorithms construct decision trees by looking at the past performance of input variables (i.e., independent variables) with respect to an outcome variable (i.e., a dependent variable). The decision tree is constructed inductively from cases with known values for the outcome variable. Input variables are sequentially selected from the data set to construct the decision tree using a divide-and-conquer algorithm that is driven by an evaluation criterion, in our case information gain. The most desirable input variable at a given point in the tree is the one with the greatest information gain. That is, the algorithm selects the input variable that requires the fewest number of subsequent splits to reach indivisibility. Using this selection process, the data are sub-divided until the set of cases is indivisible. This repeated division creates the decision tree. During the tree-construction process, branches may be pruned to increase the classification performance. Pruning simplifies the tree by replacing or removing branches that do not meet a specified confidence threshold (Witten and Frank 2005; Han and Kamber 2001).

An example of a completed classification decision tree is presented in Figure 1. This generic decision tree has four nodes or points at which decisions are made. A survey respondent would be classified by this tree as a *w*, *x*, *y*, or *z*—the outcomes or values of the dependent variable at the leaf nodes. A group of respondents (i.e., segment) with the variable values (Var1

= *a* then Var2 = *4*) would be classified as the dependent variable with outcome *x*. These respondents would be classified by following the Var1-Var2 edge of the tree and then the Var2-DepVar=*x* edge, reaching a DepVar=*x* leaf. Other groups of respondents may also reach a leaf with DepVar=*x* (e.g., Var1 = *c* then Var3 = *g* then Var4 = *s*), but via a different path thereby constituting a different segment.

<FIGURE 1 ABOUT HERE>

After the decision tree is constructed, each branch of the decision tree is converted into a rule. The tree presented in Figure 1 can be converted into the following rules:

Rule 1: IF Var1 = *a* AND Var2 ≤ 5 THEN DepVar = *x*

Rule 2: IF Var1 = *a* AND Var2 > 5 THEN DepVar = *y*

Rule 3: IF Var1 = *b* THEN DepVar = *z*

Rule 4: IF Var1 = *c* AND Var3 = *g* AND Var4 = *s* THEN DepVar = *x*

Rule 5: IF Var1 = *c* AND Var3 = *g* AND Var4 = *t* THEN DepVar = *w*

Rule 6: IF Var3 = *h* THEN DepVar = *y*

The rules provide insight into how the outcome variable's value is, in fact, dependent on the input variables, and each rule constitutes a segment of individuals. A complete decision tree provides for all possible combinations of the input variables and their allowable values reaching a single, allowable outcome. The decision tree and rules can be analyzed to predict future behavior of new data and classify segments of a market based on the outcome variable's values.

It is easy to infer incorrectly from basic data mining models that data mining is an atheoretical endeavor. For example, Roiger and Geatz (2003) present a basic data mining process in which data are collected and prepared for analysis, the prepared data are processed by the data mining software, the results are interpreted, and then the results are applied to a new problem. But effective data mining is not an atheoretical or black-box process. The domain and its associated data may contain overlapping or redundant data that require domain expertise to unravel in order to validate model performance and improve accuracy (Anand, Bell, and Hughes 1995). Or the data set may exclude important variables, which makes the generated model less useful and possibly counterproductive if used naively.

Scime and Murray (in press), Hofmann and Tierney (2003), and Ankerst, Ester, and Kriegel (2000) propose iterative processes that combine domain expertise with the data mining process. The use of the domain expert in this manner is similar to the well known use of an expert in the construction of an expert system in artificial intelligence (Giarratano and Riley 2004; Turban, McLean, and Wetherbe 2004). The intent of their iterative expert-data mining process is to use the domain expert's knowledge to increase the usefulness and accuracy of predictions and/or to reduce the number of variables required to classify the outcome variable. The process includes both iterative variable and case selection. Variable selection involves three reduction processes. First, the number of variables is reduced based on the expert's knowledge of the domain. Second, a data mining-based evaluation criterion is used to suggest a further reduction in the number of variables. Finally, the result of this reduction is reviewed by the expert for further reduction and/or enhancement based on domain knowledge. Case selection involves expert review and selection of cases that fit the goals of the data mining project such as theoretical or practical significance. The iterative variable- and case-selection process identifies the data set used for the analysis. Importantly, by data mining convention, the data set is divided into a "training set" and a "test set." The training set is composed of the data used to construct the decision tree. The test set is composed of unused, independent data that are reserved for testing the performance and reliability of the completed decision tree. This iterative process is the basis for the research presented here.

Finally, it is appropriate to note that outcomes can be predicted and interrelationships in data can be evaluated using other methods as well. For example, cluster and regression analysis can also be used to classify data. An important advantage that decision tree classification holds over cluster analysis in terms of segmentation is that decision trees classify every case in the data set, while cluster analysis does not. As well, Andoh-Baidoo and Osei-Bryson (2007) have shown that decision trees are more insightful than regression in predicting the interaction of variables on the dependent variable. More specifically, decision tree classification holds at least three significant advantages over regression analysis: (1) regression requires that missing values be estimated or that data (i.e., records or variables) be eliminated whereas decision tree algorithms maintain the integrity of the data by accounting for missing data as simply another category of a variable; (2) decision trees provide direct knowledge of how changes to the dependent variable can change the result; and, most importantly for this research, (3) a decision tree produces output that is easily converted into specific, actionable rules that, unlike regression, identify segments of like individuals.

Data and the Data Mining Method

The data mining process is most efficiently and effectively applied to databases that include extensive information on, for example, customers, patients, or, in the case of political campaigns, potential voters. Collecting this type of information is costly—for example, reports suggest that Democratic operatives planned to spend $10 million to construct an extensive voter database (Edsall 2006)—and the resulting information provides competitive advantage, so the databases are not made available to the public. Given the proprietary nature of these political data sets, in this study we rely on the most appropriate publicly availably data source for American elections, the American National Election Study (ANES). The ANES is an ongoing, long-term series of public opinion surveys intended to produce research-quality data for researchers who study the theoretical and empirical bases of American national election outcomes using voting behavior, public attitudes, and measures of political participation. Given this objective, the ANES collects data on a wide array of subjects such as voter registration and choice, social and political values, social background and structure, partisanship, candidate and group evaluations, opinions about public policy, ideological support for the political system, mass media consumption, and egalitarianism. Importantly, this collection of measures captures the types of variables (i.e., demographic, geographic, psychographic, and behavioral) that are conventionally used to segment markets (Rees and Gardner 2005). The ANES has conducted pre- and post-election, face-to-face interviews of a nationally representative sample of adults every presidential and midterm election year since 1948, except for the midterm election of 1950.

The 1948-2004 ANES Cumulative Data File is a single file composed of the pooled cases and variables from each of the studies conducted since 1948 (N=47,438). The file includes most, but not all, of the questions that have been asked in three or more ANES surveys conducted during the multi-decade time period. It is composed, therefore, of more than 900 variables, which, for comparability, have been coded in a consistent manner from year to year.

With the objective of finding variables that together are indicative of respondents' presidential vote choice (Democrat, Republican, major third party, other candidate, or abstention) and, therefore, targetable segments of voters, we applied the iterative expert-data mining methodology—the C4.5 classification algorithm and domain expertise—to the 1948-

2004 ANES Cumulative Data File to construct a decision tree and resulting rules. We analyzed the multi-decade series instead of focusing on the most recent election year(s) in order to capture long-term predictors that have persisted, and are more likely to persist, over time. Computing resources—both software and hardware—dictated that we reduce the number of variables in the data set from the more than 900 available in the ANES data file to 250 or less. Using domain expertise, primarily derived from scholarly literature (e.g., Lacy and Burden 1999, Alvarez and Nagler 1995), we reduced the data set to 237 variables in addition to the outcome or dependent variable. We excluded variables that appeared to be variations of the outcome variable that would be predictive mostly due to their high correlation with the outcome variable. This class of variables included measures such as the configuration of a respondent's split ticket vote where both the party of presidential and congressional vote were identified. Of the remaining variables, we retained those that held the most empirical and theoretical significance. The list of variables is available from the authors upon request.

In terms of case selection, the data were limited to respondents to surveys from presidential election years, thereby excluding midterm election years. This selection is based on the recognition that more extensive surveys were administered during presidential years as well as the increased attention and interest paid by citizens during these high-profile elections. The resulting data set includes more than 26,000 presidential-year records from which we randomly selected a subset of 6678 respondents.[1] The data set of 237 variables and 6678 respondents, then, composes the training data set from which the decision trees are constructed. The remaining presidential-year data were retained as the test set (20,033 records).

[1] A test of the selected cases indicates that the randomization process was successful.

Next, we estimated the information gain of each of the variables in the training set and ordered them by information gain from highest to lowest.[2,3] Then, we constructed a series of decision trees using the WEKA implementation of the classification decision tree C4.5 algorithm with three-fold cross-validation (Witten and Frank 2005). That is, we executed the classification algorithm repeatedly, with each execution producing a decision tree from progressively smaller sets of variables. The first execution used all 237 variables, with the lowest information-gain-ranked variables removed from each succeeding execution until the final execution, which was composed of the nine variables exhibiting the highest information gain. We then compared all the resulting decision trees for predictive accuracy (i.e., the measure of the number of instances correctly classified by the derived decision tree) to determine the most accurate tree to use to produce rules. Since a rule represents a segment, more accurate predictions by decision trees indicate more accurate rules and, therefore, more reliable segmentation.

In the final stage, we reviewed the tree and used domain knowledge for further variable additions and deletions to the data set. It is possible, for example, for data mining output to be influenced by the order in which records are processed. This can affect the information gain calculation, which in turn can cause a critical variable either to be included or excluded from the final decision tree. This review process enables the domain expert to add or delete critical variables that are likely to increase the tree's accuracy and/or usefulness. It is also important to note that following data mining convention the classification decision trees were constructed and evaluated using the training data set. When the domain expert was satisfied with the classification tree, it was tested for validity with the test data set of 20,033 records.

[2] Information gain was estimated using WEKA (Witten and Frank 2005).
[3] This is just one approach to selecting independent variables to model. For example, Roiger and Geatz (2003) propose using unsupervised cluster analysis, another data mining technique, to select variables.

Results: The Classification Decision Tree

In this section we review the results of the decision tree classification process. These results include (1) identification of the set of variables that construct the decision tree manifesting the greatest predictive power before any theory-driven adjustments by the domain expert, (2) an analysis of theory-driven adjustments to the set of variables by the domain expert, and (3) an analysis of the structure of the final decision tree to identify the rules and, therefore, the segments of the electorate it provides. The first column of Table 1 indicates the set of variables with the greatest predictive power before the domain expert adjusted the variable list. Columns 2 and 3 identify a number of domain-expert adjustments of that variable set, which were used to construct the final decision tree. Column 4 indicates the variables appearing in the final decision tree and, therefore, the rules.

<TABLE 1 ABOUT HERE>

The first step in the analytical process was to construct the most accurate decision tree using the C4.5 algorithm without theory-driven adjustment by the domain expert.[4] Set 1 includes 20 variables (the outcome variable and 19 input variables) with a 64.2% success rate of correctly predicting the vote choice (i.e., Democrat, Republican, major third party, other candidate, or abstention). This list of variables clearly speaks to the literature regarding the importance of campaigns. On one hand, the list includes variables that suggest that voting choice is beyond the reach of campaigns. For example, among the list of predetermined measures are education, party identification, and interest in public affairs. On the other hand, though, there are a number of variables that indicate the importance of campaigns. These include affect toward the Democratic and Republican presidential and vice presidential candidates and beliefs about the party most

[4] The decision trees are not reported here due to space constraints. For example, the final decision tree has 330 nodes. The trees are available from the authors upon request.

likely to address the respondent’s most important national problem. Notably, the root node or first decision point in the tree is affect for the Republican presidential candidate, which is followed at the first child node by affect for the Democratic presidential candidate. In other words, the two most salient variables for vote choice are voter attitudes that are subject to campaign manipulation such as political advertising (e.g., Kaid and Chanslor 2004). Other variables that may or may not be affected by campaigns include affect and feeling thermometers toward the Democratic and Republican Parties, interest in the election, and feeling thermometers toward conservatives and liberals. In all, the initial results of this data mining process, which captured and analyzed a small set of predictive variables (19) from a large number of potential variables (>900), suggest that the campaigns do indeed matter. And they suggest that one of the primary influences is through affect toward the candidates.

The next step in the analytical process involved applying theory-driven domain expertise to evaluate and to adjust the set of variables. Among these variables is a question asking for whom respondents voted in the corresponding congressional election. We identified this as a post-election question that is inappropriate to retain since the intent is to predict voter behavior prior to the election. Consistent with our process, we removed this variable from the set of variables. This resulted in an inconsequential change in the success rate of the resulting decision tree.

Further review of the variables also led to the decision to evaluate other variables in an effort to improve the performance and usefulness of the decision tree. For example, race, sex, marital status, religion, and aggregate indicators of economic assessments were not included in the set of variables that produced the most successful pre-expert-adjustment decision tree. Race (e.g., Burton and Shea 2003), sex (e.g., Seltzer, Newman, and Leighton 1997), marital status (e.g., Weisberg 1987), and religion (e.g., Leege and Kellstedt 1993) are considered strongly indicative of voting behavior, each holding both theoretical and targeting significance. There is also strong evidence of a relationship between economic performance and vote choice (Abramson, Aldrich, and Rohde 2003). While there are indications of a strong retrospective component to the voting decision (Fiorina 1981), there are also indications that voters engage in prospective voting (MacKuen, Erikson, and Stimson 1992). Measures of race, sex, marital status, and religion were forced back into the set of variables for further analysis. We also forced back into the data set aggregate evaluations of retrospective and prospective economic performance, following Nadeau and Lewis-Beck (2001).[5]

We also evaluated other variables to force back in to the data set. The variables tested were those that would be expected to improve further the model's accuracy. These include variables pertaining to beliefs about traditional values and religion; feeling thermometers about immigrants, the economy, blacks, gays, welfare, and abortion; residence in the South; voting in the primary election; and tolerance of different moral standards. The results indicated that a highly accurate decision tree (65.2% success rate) is constructed from the set of variables that includes Set 1 minus congressional vote plus the expert-identified variables for race, the assessment of whether all people receive an equal chance, and feelings toward abortion and gays. These variables constitute Set 2 in Table 1. It is worth noting that adding just race to Set 1 increases the success rate to 64.8% (not shown in Table 1).

Following data mining convention, we then deleted each individual variable from Set 2 and tested the effect of each deletion on the accuracy of the resulting decision trees. The variables that when deleted improved the ability to predict vote by increasing the percentage of

[5] We enhanced the ANES cumulative data set with these aggregate evaluations of economic performance, which are the only non-ANES variables used in the analysis.

correctly classified instances were: concern for war, the actual closeness of the election, the political party handling the most important problem, and feeling thermometers about the Democratic vice presidential candidate and conservatives. Deleting these variables as a group provides a set of 17 variables that generates our final and most accurate decision tree with a success rate of 65.6%. This set of variables constitutes Set 3. In order to give some context, this 65.6% success rate may be compared to Lacy and Burden (1999) who produced a multinomial probit model, which also included abstention as a vote choice, that correctly classified 50.6% of respondents. On the other hand, Alvarez and Nagler (1995) produced a multinomial probit model that correctly classified 74.0% of respondents but did not attempt to predict abstention as a vote choice. This comparison suggests that decision tree analysis produces accuracy rates that compare favorably to other methods of prediction. It is important to note as well that there is evidence that decision trees enjoy an advantage over regression analysis in terms of the ease of application of the output, particularly for the purposes of segmentation, and the handling of missing values (Andoh-Baidoo and Osei-Bryson 2007).

Of course, we recognize that our final decision tree may be specific to the data from which it was constructed and, therefore, lack external validity. In order to test its generalizability and robustness, we used the final decision tree to classify the test data that we reserved for this purpose. These data include the same variables as the training data, but are composed of cases that were not used to build the model. That is, the test set is composed of the remaining 20,033 records in the presidential-year data. For the test set, the tree correctly predicted vote choice in 64.4% of the cases (versus 65.6% for the training set). Based on this result, we conclude that the decision tree is externally valid and the resulting rules/segments are highly reliable.

Results: The Rules (i.e., Segments)

The final analytical step is to write the rules that are provided by the decision tree. A rule is a heuristic that can be used to predict how individuals with a specific set of common characteristics will behave. In particular, each rule represents a segment. That is, rules are useful for segmenting markets and, more importantly here, for targeting appeals to voters. To do this, we examined the final decision tree, which results from the Set 3 variables. During this examination, we noted that the C4.5 algorithm obviated the need for four of the final 17 variables (abortion, unequal chances, and feeling thermometers about gays and liberals) while obtaining the same accuracy at 65.6%. This means that these four variables are needed to build the more accurate decision tree, but are not needed for future prediction. When there are insufficient instances of the data to develop a branch, as determined by an assigned threshold, the C4.5 algorithm prunes the tree. That is, it removes the variables composing that segment of the branch and rolls the effected instances into the preceding node and then into other appropriate branches. This pruning process can (and in our case does) reduce the number of variables in the tree. As a result, there are 13 variables that are used in the resulting classification rules.

The final decision tree can be converted into 223 rules or segments. Each rule represents one branch of the decision tree from root node to a leaf node. Clearly not all variables occur on every branch (see Figure 1). However, taken together these 223 rules account for all the variables in the training set and will always result in a determination of voting choice. For example, Rule 12 indicates that respondents with the following characteristics will abstain 66.4% of the time:

> Rule 12: IF affect toward Republican presidential candidate is neutral or slightly negative (0 or -1) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is weak Democrat AND interest in campaign is “not much”

AND feeling for Democratic Party is positive (>55) AND affect toward the Republican Party is slightly negative to positive (>-3)

THEN individuals ABSTAIN.

Importantly, though, Rule 11 indicates that respondents with these same characteristics except for greater negative affect toward the Republican candidate (<-2 versus 0 or -1) will vote for the Democratic presidential candidate 49.4% of the time. This suggests, among other things, the potential to convert non-voters to Democratic voters through strategic considerations related to campaign message and tone such as negative or comparative advertising (e.g., Fridkin and Kenney [2004] found that negative messages containing relevant information and delivered by non-controversial means can depress evaluations of opponents).

Given the resource constraints on most campaigns, rules should be evaluated in terms of their accuracy as well as the proportion of cases to which they apply; that is, segments should be assessed in terms of their efficiency. For example, a rule that is 100% accurate but applies to only a handful of cases must be evaluated relative to a rule that is 60% accurate but applies to a large proportion of the cases. In this instance, 15 rules apply to 1% or more of the cases (N ≥67), with the most cases associated with Rule 203, which applies to 855 respondents, and the fewest cases associated with Rule 134, which applies to 86 respondents. On the other hand, 61 rules correctly predict voting choice in at least 90% of the applicable cases. These 61 rules, though, apply to only three respondents on average and, therefore, probably hold little value for segmentation or targeting.

With these substantive issues in mind, Table 2 presents summary information on the 15 rules that apply to 1% or more of the cases. These 15 rules are presented in Appendix B. Together, they account for well more than half (61.6%) of the sample. The mean accuracy of the

15 rules is 65.8%, with a high of 86.9% (Rule 223) and a low of 45.2% (Rule 173). These rules employ 11 of the 13 variables of the variable set. The unused variables (feeling thermometer about the Republican vice presidential candidate and education) may be dropped from consideration now that the data mining is complete. The 15 rules along with their complementary rules can be used to strategically plan campaigns.

<TABLE 2 ABOUT HERE>

As one would expect, some rules are more useful than others. For example, Rule 203 is the rule that applies to the largest proportion of the electorate. While it is consistent with what most knowledgeable observers would expect and while it supports the validity of the data mining process, it is not strategically useful or informative. It states:

> Rule 203: IF affect toward Republican presidential candidate is positive (>0) AND party identification is weak or leaning Republican AND race is white
>
> THEN individuals vote REPUBLICAN.

At the same time, a number of rules, such as Rules 11 and 12 discussed above, hold theoretical and strategic interest. For another example, rule 200 addresses a segment of independent identifiers:

> Rule 200: IF affect toward Republican presidential candidate is positive (>0) AND affect toward Democratic presidential candidate is positive to negative (<3) AND party identification is independent AND interest in the campaign is “somewhat” AND affect toward the Democratic Party is negative to positive (>-3) AND affect toward the Republican Party is negative to positive (>-3)
>
> THEN individuals vote REPUBLICAN.

Importantly, though, Rule 200 implies a means to convert these independents from voting Republican to voting Democratic. When evaluated in conjunction with Rule 199, which is not one of the 15 primary rules due to the limited number of individuals to which it applies but is useful for the analysis of a rule that is, Rule 200 suggests that independent respondents with these same characteristics except for greater negative affect toward the Republican Party (≤-3 versus >-3) will vote for the Democratic presidential candidate 47.3% of the time. Rules 199 and 200 taken together suggest the potential to convert a Republican voter to a Democratic voter through strategic campaign efforts such as increasing negative affect toward the Republican Party through negative or comparative advertising that criticizes, for example, the party's management of the economy or international relations.

In a similar vein, Rule 9 addresses a segment of non-voters. It states:

> Rule 9: IF affect toward Republican presidential candidate is neutral (0) AND affect toward Democratic presidential candidate is neutral (0) AND party identification is weak Democrat AND feeling toward the Democratic Party is slightly warm to cold (≤55)
>
> THEN individuals ABSTAIN.

Importantly, though, Rule 8, which is not one of the 15 primary rules but is useful for the analysis of Rule 9, indicates that respondents with these same characteristics except for greater negative affect toward the Democratic candidate (<-1 versus 0) will vote for the Republican presidential candidate 48.1% of the time. This suggests the potential to convert non-voters to Republican voters through strategic considerations related to campaign message and tone such as increasing negative affect toward the Democratic candidate by noting, for example, perceived flaws in the candidate's personal or professional life.

In all, this brief review of a few of the rules demonstrates some of the analysis that is possible with decision tree classification. Further, the rules suggest that campaign message and tone may be adjusted to influence certain groups or segments of voters. Importantly, decision tree analysis identifies characteristic by characteristic who these voters are.

Conclusion

We data mined the 900-item ANES to reduce dramatically the number of variables needed to predict voting behavior and, more importantly, to segment the electorate in useful and meaningful ways. The results demonstrate the potential for an emerging application in political marketing. The data mining methodology not only indicates that campaigns matter, but also presents usable outputs that provide a mechanism for them to matter more. Of course, this small set of variables presented in the results could be used in conjunction with opinion surveys to predict voting choice and behavior. In all, the findings include a number of useful rules that allow micro-targeting through the successful segmentation of the electorate by accurately predicting almost two-thirds of the voting decisions.

The interesting results highlight the possibilities for future study. In this research we attempted to improve the accuracy of the predictions of vote choice, a process that by definition significantly narrowed the number of variables employed. This, of course, has ramifications for the rules that are provided by the decision tree and, more importantly, the associations that are detected. In subsequent research we intend to cast a wider net to look for associations in the complete set of potential variables as well as to analyze other data sets. Further, we believe the data mining process may shed new light on other important outcome variables such as turnout specifically and other forms of participation.

Finally, we believe these results reconfirm the proposition that campaigns matter, and they identify a mechanism through which they can matter more. Not only do these findings suggest that the data mining methodology may increase efficiency for political campaigns, but they also suggest that, from a democratic theory perspective, overall participation may be improved by identifying and communicating more effective messages that better inform intended voters and that motivate individuals to vote who otherwise may abstain.

Appendix A: ANES Survey Items

Outcome Variable

Summary variable indicating the respondent's presidential vote choice or abstention. ANES data set variable VCF0706.

1. Democrat
2. Republican
3. Major third party
4. Other
7. Did not vote or voted but not for president

Input Variables

Abortion. "There has been some discussion about abortion during recent years. Which one of the opinions on this page best agrees with your view?" ANES data set variable VCF0838.

1. By law, abortion should never be permitted.
2. The law should permit abortion only in case of rape, incest, or when the woman's life is in danger.
3. The law should permit abortion for reasons other than rape, incest, or danger to the woman's life, but only after the need for the abortion has been clearly established.
4. By law, a woman should always be able to obtain an abortion as a matter of personal choice.

Affect toward the Democratic Party. This is the number of Democratic Party "likes" mentioned by the respondent minus the number of Democratic Party "dislikes" mentioned. ANES data set variable VCF0316.

Affect toward Democratic presidential candidate. This is the number of Democratic presidential candidate "likes" mentioned by the respondent minus the number of Democratic presidential candidate "dislikes" mentioned. ANES data set variable VCF0403.

Affect toward Republican Party. This is the number of Republican Party "likes" mentioned by the respondent minus the number of Republican Party "dislikes" mentioned. ANES data set variable VCF0320.

Affect toward Republican presidential candidate. This is the number of Republican presidential candidate "likes" mentioned by the respondent minus the number of Republican presidential candidate "dislikes" mentioned. ANES data set variable VCF0407.

Avoid war. "Looking ahead, do you think the problem of keeping out of war would be handled better in the next four years by the Republicans, by the Democrats, or about the same by both?" ANES data set variable VCF0522.

1. Better by Democrats
2. Same by both
3. Better by Republicans
9. Don't know/Depends/Neither

Close election. Difference in percent of the vote received by the top two candidates in the presidential election. Data for 1948-2000 were collected from http://presidentelect.org. Data for 2004 were collected from http://www.cnn.com/ELECTION/2004/.

Congressional vote. "How about the election for the House of Representatives in Washington? Did you vote for a candidate for the U.S. House of Representatives?" If yes, "Who did you vote for?" ANES data set variable VCF0707.

1. Democrat
2. Republican

Conservatives—feeling thermometer. See Democratic presidential candidate feeling thermometer for question wording. ANES data set variable VCF0212.

00-96. Degrees as coded
97. 97-100 Degrees

Democratic presidential candidate—feeling thermometer (DEMTHERM). "I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person and I'd like you to rate that person using something we call the feeling thermometer. Ratings between 50 and 100 degrees mean that you feel favorably and warm toward the person; ratings between 0 and 50 degrees mean that you don't feel favorably toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person. If we come to a person whose name you don't recognize, you don't need to rate that person. Just tell me and we'll move on to the next one." ANES data set variable VCF0424.

00-96. Degrees as coded
97. 97-100 Degrees

Democratic vice presidential candidate—feeling thermometer. See DEMTHERM for question wording. ANES data set variable VCF0425.

00-96. Degrees as coded
97. 97-100 Degrees

Economic Future Index: Prospective Economic Evaluation. "Now turning to business conditions in the country as a whole—do you think that during the next 12 months we'll have good times financially, or bad times or what?” Following Nadeau and Lewis-Beck (2001), the score is the percentage of respondents saying “good times” minus percentage saying “bad times” for the last quarter of every presidential year. This item is part of the Surveys of Consumer Attitudes and Behavior, University of Michigan. November is the month of interest. For 1956 to 1976, the November results are used. For 1980 to 2000, the November results are averaged with the October and December results.

Education. What is the highest degree that you have earned? ANES data set variable VCF0140a.

1. 8 grades or less
2. 9-12 grades, no diploma/equivalency
3. 12 grades, diploma or equivalency
4. 12 grades, diploma or equivalency plus non-academic training
5. Some college, no degree; junior/community college level degree (AA degree)
6. BA level degrees
7. Advanced degrees including LLB

Gay men and lesbians; i.e., homosexuals—feeling thermometer. See DEMTHERM for question wording. ANES data set variable VCF0232.

00-96. Degrees as coded
97. 97-100 Degrees

Interest in election. “Would you say that you have been/were very much interested, somewhat interested, or not much interested in the political campaigns this year?” ANES data set variable VCF0310.

1. Not much interested
2. Somewhat interested
3. Very much interested

Interest in public affairs. “Would you say you follow what’s going on in government and public affairs most of the time, some of the time, only now and then, or hardly at all?” ANES data set variable VCF0313.

1. Hardly at all
2. Only now and then
3. Some of the time
4. Most of the time

Liberals—feeling thermometer. See DEMTHERM for question wording. ANES data set variable VCF0211.

00-96. Degrees as coded
97. 97-100 Degrees

National Business Index: Retrospective Economic Evaluation. “Would you say that at the present time business conditions are better or worse than they were a year ago?” Following Nadeau and Lewis-Beck (2001), the score is the percentage of respondents saying “better” minus percentage saying “worse” for the last quarter of every presidential year. This item is part of the Surveys of Consumer Attitudes and Behavior, University of Michigan. November is the month of interest. For 1956 to 1976, the November results are used. For 1980 to 2000, the November results are averaged with the October and December results.

Party identification. Summary measure of party identification. ANES data set variable VCF0301.

-3 strong Republican
-2 weak or leaning Republican
0 Independent
2 weak or leaning Democrat
3 strong Democrat

Party problem. If the respondent mentioned a "most important problem" in the nation in an earlier question, "Which political party do you think would be most likely to get the government to do a better job in dealing with this problem—the Republicans, the Democrats, or wouldn't there be much difference between them?" ANES data set variable VCF9012.

1. Democrats
3. Not much difference
5. Republicans

Race. "In addition to being American, what do you consider your main ethnic group or nationality group?" ANES data set variable VCF0106a.

1. White
2. Black
3. Asian
4. Native American
5. Hispanic
7. Other

Republican presidential candidate—feeling thermometer. See DEMTHERM for question wording. ANES data set variable VCF0426.

00-96. Degrees as coded
97. 97-100 Degrees

Republican vice presidential candidate—feeling thermometer. See DEMTHERM for question wording. ANES data set variable VCF0427.

00-96. Degrees as coded
97. 97-100 Degrees

Unequal chances. "One of the big problems in this country is that we don't give everyone an equal chance." ANES data set variable VCF9015.

1. Agree strongly
2. Agree somewhat
3. Neither agree nor disagree
4. Disagree somewhat
5. Disagree strongly

Who elected. “Who do you think will be elected President in November?” ANES data set variable VCF0700.

1. Democratic candidate
2. Republican candidate
7. Other candidate

Appendix B: Rules

Rule 9: IF affect toward Republican presidential candidate is neutral (0) AND affect toward Democratic presidential candidate is neutral (0) AND party identification is weak Democrat AND feeling toward the Democratic Party is neutral to cold (≤55)

THEN individuals ABSTAIN.

Rule 12: IF affect toward Republican presidential candidate is neutral or slightly negative (0 or -1) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is weak Democrat AND interest in campaign is “not much” AND feeling for Democratic Party is warm (>55) AND affect toward the Republican Party is slightly negative to positive (>-3)

THEN individuals ABSTAIN.

Rule 13: IF affect toward Republican presidential candidate is neutral or negative (≤0) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is weak or leaning Democrat AND feeling for Democratic Party is warm (>55) AND interest in campaign is “somewhat”

THEN individuals vote DEMOCRAT

Rule 14: IF affect toward Republican presidential candidate is neutral or negative (≤0) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is weak or leaning Democrat AND feeling for Democratic Party is warm (>55) AND interest in campaign is “very much”

THEN individuals vote DEMOCRAT

Rule 35: IF affect toward Republican presidential candidate is neutral or negative (≤0) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is strong Democrat AND affect toward Democratic Party is positive (>0)

THEN individuals vote DEMOCRAT

Rule 36: IF affect toward Republican presidential candidate is neutral or negative (≤0) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is independent AND interest in campaigns is “not much”

THEN individuals ABSTAIN

Rule 61: IF affect toward Republican presidential candidate is neutral or negative (≤0) AND affect toward Democratic presidential candidate is neutral or negative (≤0) AND party identification is weak or leaning Republican AND race is white AND feeling for Republican Party is warm (>55) AND follows public affairs “some of the time”

THEN individuals vote REPUBLICAN

Rule 80: IF affect toward Republican presidential candidate is neutral to negative (≤0) AND affect toward Democratic presidential candidate is positive (>0) AND party identification is weak or leaning Democrat AND race is white

THEN individuals vote DEMOCRAT

Rule 107: IF affect toward Republican presidential candidate is neutral to negative (≤0) AND affect toward Democratic presidential candidate is positive (>0) AND party identification is strong Democrat

THEN individuals vote DEMOCRAT

Rule 134: IF Affect toward Republican presidential candidate is neutral to negative (≤0) AND affect toward the Democratic presidential candidate is positive (>0) AND party identification is weak Republican AND respondent believes the Democratic candidate will win AND race is white

THEN individuals vote DEMOCRAT.

Rule 154: IF affect toward Republican presidential candidate is positive (>0) AND party identification is weak or leaning Democrat AND feeling for Democratic Party is slightly warm to cold (≤65) AND affect toward Democratic presidential candidate negative (≤-1) AND interest in the campaign is “somewhat”

THEN individuals vote REPUBLICAN

Rule 173: IF affect toward Republican presidential candidate is positive (>0) AND party identification is weak or leaning Democrat AND feeling for Democratic Party is warm (>65)

THEN individuals vote DEMOCRAT

Rule 200: IF affect toward Republican presidential candidate is positive (>0) AND party identification is independent AND affect toward Democratic presidential candidate is slightly positive to negative (≤3) AND affect toward Democratic Party is slightly negative to positive (>-3) AND interest in campaign is “somewhat” AND affect toward Republican Party is slightly negative to positive (>-3)

THEN individuals vote REPUBLICAN

Rule 203: IF affect toward Republican presidential candidate is positive (>0) AND party identification is weak or leaning Republican AND race is white

THEN individuals vote REPUBLICAN

Rule 223: IF affect toward Republican presidential candidate is positive (>0) AND party identification is strong Republican

THEN individuals vote REPUBLICAN

References

Alvarez, R. Michael and Jonathan Nagler. 1995. “Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election.” *American Journal of Political Science,* 39, 3: 714-44.

American National Election Studies. 2005. Center for Political Studies, University of Michigan, Ann Arbor, MI.

Anand, Sarabjot. S., David. A. Bell and John. G. Hughes. 1995. “The Role of Domain Knowledge in Data Mining.” *Proceedings of the Fourth International Conference on Information and Knowledge Management*, 37-43. Baltimore, MD.

Andoh-Baidoo, Francis K. and Kweku-Muata Osei-Bryson. 2007. “Exploring the Characteristics of Internet Security Breaches that Impact the Market Value of Breached Firms.” *Expert Systems with Applications*, 32, 3: 703-25.

Ankerst, Mihael, Martin Ester and Hans-Peter Kriegel. 2000. “Towards an Effective Cooperation of the User and the Computer for Classification.” *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 179-88. Boston, MA.

Burton, Michael J. and Daniel M. Shea. 2003. *Campaign Mode: Strategic Vision in Congressional Elections*. New York: Rowman and Littlefield.

Deshpande, Mukund and George Karypis. 2002. “Using Conjunction of Attribute Values for Classification.” *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, November 4-9: 356-64. McLean, VA.

DuMouchel, William and Daryl Pregibon. 2001. “Empirical Bayes Screening for Multi-item Associations.” *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, 67-76.

Edsall, Thomas B. 2006. “Democrats’ Data Mining Stirs an Intraparty Battle.” *The Washington Post*, March 8: A1.

Freitas, Alex. A. 2000. “Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A Position Paper.” *ACM SIGKDD Explorations Newsletter*, 2, 1: 65-69.

Fridkin, Kim Leslie and Patrick J. Kenney. 2004. “Do Negative Messages Work? The Impact of Negativity on Citizens’ Evaluations of Candidates.” *American Politics Research,* 32, 5: 570-605.

Giarratano, Joseph. C. and Gary. D Riley. 2004. *Expert Systems: Principles and Programming,* 4th Edition. New York: Course Technology.

Hamburger, Tom and Peter Wallsten. 2005. “Parties Are Tracking Your Habits: Though both Democrats and Republicans collect personal information, the GOP’s mastery of data is changing the very nature of campaigning.” *LA Times* July 24. Available at latimes.com Accessed July 24, 2005.

Han, Jiawei and Micheline Kamber. 2001. *Data Mining: Concepts and Techniques*. Boston: Morgan Kaufmann.

Hofmann, Markus and Brendan Tierney. 2003. “The Involvement of Human Resources in Large Scale Data Mining Projects.” *Proceedings of the 1st International Symposium on Information and Communication Technologies,* 103-109. Dublin, Ireland.

Jaroszewicz, Szymon and Dan A. Simovici. 2004. “Interestingness of Frequent Itemsets Using Bayesian Networks as Background Knowledge.” *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* August 22 – 25. Seattle, WA.

Kaid, Lynda Lee and Mike Chanslor. 2004. “The Effects of Political Advertising on Candidate Images.” In *Presidential Candidate Images*, ed. Kenneth L. Hacker. Westport, CT: Praeger.

Lacy, Dean and Barry C. Burden. 1999. “The Vote-Stealing and Turnout Effects of Ross Perot in the 1992 U.S. Presidential Election.” *American Journal of Political Science*, 43, 1: 233-55.

Leege, David C. and Lyman Kellstedt. 1993. *Rediscovering the Religious Factor in American Politics.* Armonk, NY: M. E. Sharpe.

McCarthy, V. 1997. “Strike It Rich” *Datamation,* 43, 2: 44-50.

Nadeau, Richard and Michael S. Lewis-Beck. 2001. “National Economic Voting in U.S. Presidential Elections.” *Journal of Politics* 63, 1: 159-181.

Niemi, Richard G. and Herbert F. Weisberg. 2001. “What determines the vote?” In *Controversies in Voting Behavior*, 4th edition, ed. R. G. Niemi and H. F. Weisberg. Washington, DC: CQ Press.

Osei-Bryson, Kweku-Muata. 2004. “Evaluation of Decision Trees: A Multicriteria Approach” *Computers and Operations Research*, 31, 11: 1933-1945.

Padmanabhan, Balaji and Alexander Tuzhilin. 2000. “Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns.” *Proceedings of the Sixth ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 54-63. Boston, MA.

Rees, Patricia and Hanne Gardner. 2005. “Political Marketing Segmentation – The Case of UK Local Government.” In *Current Issues in Political Marketing*, eds. Walter W. Wymer and Jennifer Lees-Marshment. Binghamton, NY: Best Business Books.

Roiger, Richard J. and Michael W. Geatz. 2003. *Data Mining: A Tutorial-Based Primer*. New York: Addison Wesley.

Scholz, Martin. 2005. “Sampling-Based Sequential Subgroup Mining.” *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 21-24. Chicago, IL.

Scime, Anthony and Gregg R. Murray. In press. “Vote Prediction by Iterative Domain Knowledge and Attribute Elimination.” *International Journal of Business Intelligence and Data Mining*.

Seltzer, Richard A., Jody Newman and Melissa Voorhees Leighton. 1997. *Sex as a Political Variable: Women as Candidates and Voters in U.S. Elections* Boulder, CO: Lynne Rienner.

Shea, Daniel M. and Michael John Burton. 2006. *Campaign Craft: The Strategies, Tactics, and Art of Political Campaign Management*, 3rd edition. Westport, CT: Praeger.

Turban, Efraim, Ephraim McLean and James Wetherbe. 2004. I*nformation Technology for Management*, 3rd edition. New York: Wiley.

Weisberg, Herbert F. 1987. “The Demographics of a New Voting Gap: Marital Differences in American Voting.” *Public Opinion Quarterly,* 51: 335-43.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. San Francisco, CA: Morgan Kaufmann.

TABLE 1 Data Set Variables Used to Classify Vote Choice: By Information Gain and Expert Identification

	Set 1	Set 2	Set 3	223 Rule Set	15 Rule Set
Information Gain Identified					
Party identification	+	+	+	+	+
Affect: Rep candidate	+	+	+	+	+
Affect: Dem candidate	+	+	+	+	+
Affect: Rep Party	+	+	+	+	+
Congressional vote	+	-	-	-	-
Affect: Dem Party	+	+	+	+	+
Thermometer: Rep Party	+	+	+	+	+
Thermometer: Dem VP candidate	+	+	-	-	-
Party avoid war	+	+	-	-	-
Who will be elected	+	+	+	+	+
Party handle important problem	+	+	-	-	-
Thermometer: Rep VP candidate	+	+	+	+	-
Interest in election	+	+	+	+	+
Thermometer: Dem Party	+	+	+	+	+
Thermometer: Liberals	+	+	+	-	-
Thermometer: Conservatives	+	+	-	-	-
Follow public affairs	+	+	+	+	+
Close election	+	+	-	-	-
Education	+	+	+	+	-
Expert Identified					
Abortion	-	+	+	-	-
Race	-	+	+	+	+
Thermometer: Gays	-	+	+	-	-
Unequal treatment	-	+	+	-	-
Variables (N)	19	22	17	13	13
Correctly Predicted (%)	64.2	65.2	65.6	65.6	

TABLE 2 Decision Tree Rules: Cases (N) and accuracy

RULE#	*N*	*Accuracy (%)*
203	855	72.3
107	761	77.9
80	648	69.1
223	597	86.9
13	186	56.0
173	182	45.2
35	119	69.7
36	110	72.1
9	105	61.4
12	97	66.4
154	94	64.7
200	94	57.1
61	91	66.7
14	90	66.6
134	86	55.2

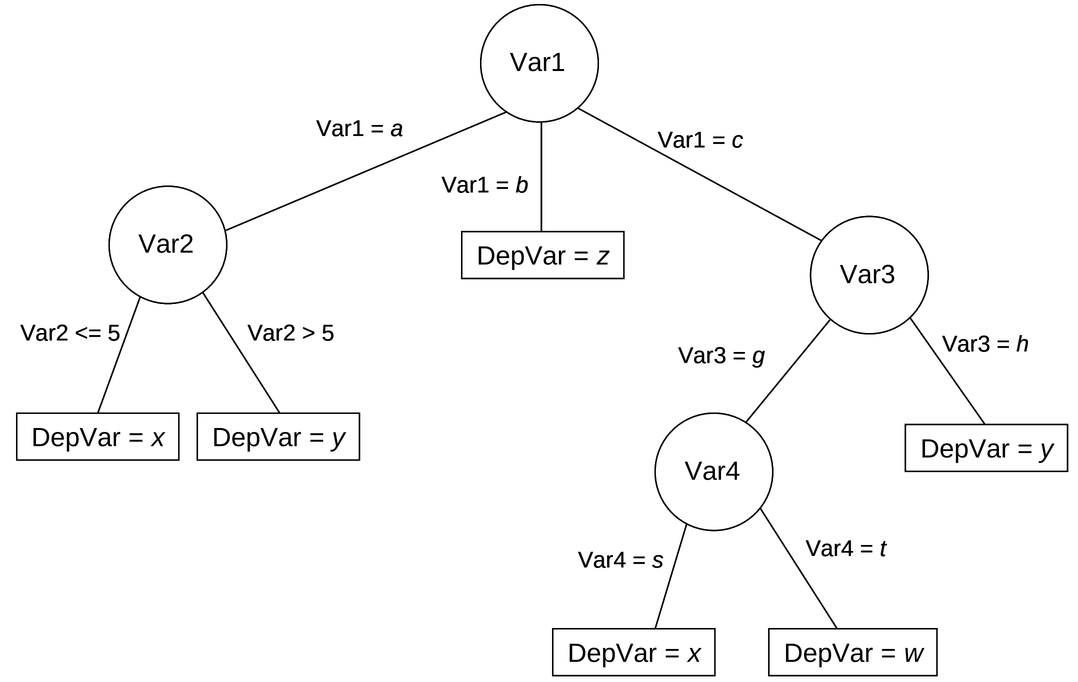

FIGURE 1 Generic Decision Tree